Standardization of Small Lesion Contrast in PET Imaging

by

Drake Cole Brookins

Department of Medical Physics
Duke University

Date:_____

Approved:

_____
Timothy Turkington, Supervisor

_____
Bennett Chin

_____
James Colsher

Thesis submitted in partial fulfillment of
the requirements for the degree of
Master of Science in the Graduate Program of
Medical Physics in the Graduate School
of Duke University

2014

ABSTRACT

Standardization of Small Lesion Contrast in PET Imaging

by

Drake Cole Brookins

Department of Medical Physics
Duke University

Date:_____
Approved:

_____
Timothy Turkington, Supervisor

_____
Bennett Chin

_____
James Colsher

An abstract of a thesis submitted in partial
fulfillment of the requirements for the degree
of Master of Science in the Graduate Program of
Medical Physics in the Graduate School of
Duke University

2014

Abstract

Quantitative measurements in PET imaging have recently become more widespread as a way to diagnose and stage many types of malignant cancer. Currently patients need to have follow-up scans performed on the same PET system due to technical factors. Multi-clinic studies using quantitative PET measurements are also confounded by these technological factors. This work aims to evaluate the use of commonly available phantoms to cross-calibrate processing parameters to equalize small lesion quantitation. The method was verified using an abdomen phantom with small hot sphere inserts, as well as a smaller phantom with small hot sphere inserts.

Methods: A GE Discovery 690 and STE were used. Both time-of-flight (TOF) and non-TOF images were used from the D690. Jaszczak phantoms with hot rod and cold rod inserts were scanned on both systems consecutively for 20 minutes. Images were reconstructed with a range of iterations and post-smoothed (PS) with 2-10 mm of smoothing. Automated analysis of the images used the CT images to find rods and then calculate a rod to background ratio for each rod sector, PET image variant, and scanner. A target rod contrast could then be chosen and parameters determined for both systems separately to equalize rod contrast. Iteration-based resolution control and PS were both evaluated. To verify, an abdomen phantom was filled with a low background activity and ten 10-mm diameter spheres filled with FDG and CT contrast. In order to evaluate

any size dependence, six 10-mm diameter spheres filled with FDG and CT contrast were placed inside a Jaszczak container filled with low background activity. An automatic CT-based analysis of the spheres was performed, obtaining mean and maximum values across the spheres.

Results: Small sphere quantitation differed substantially for similar processing between systems. However, sphere quantitation matched well when cross-calibrating the DSTE and non-TOF D690 Jaszczak phantom images by independently limiting iterations. Doing the same process with post-smoothing yielded similar results, with high iteration PS performing slightly better than PS at iterations used clinically at Duke for twenty-minute scans. Equalizing TOF images from the D690 with DSTE images with spheres placed in an abdomen phantom resulted in relatively poor correlation, but correlated well with spheres placed inside the Jaszczak phantom. Shorter scan durations behaved similarly to the twenty-minute scans.

Conclusions: Both Jaszczak phantoms worked well for cross-calibrating processing parameters to equalize quantitation in small lesions for non-TOF imaging. Iterations and PS could both be used to control resolution. It appears the best method is to use PS to fine-tune the resolution. The size dependence of TOF, and PET in general, seems to be an issue.

Contents

List of Tables

List of Figures

Acknowledgements

I'd like to thank Dr. Turkington for being a great advisor. You went above and beyond what I expected with your patience, enthusiasm, and humor. I'd also like to thank Robin Davis for always being willing to lend a hand, and all the technologists and staff involved with the PET department at Duke. I thank Josh Wilson for being a great resource of knowledge, as well as for graciously providing programs and data. Shawn Murphy and Michael Daley in the cyclotron workshop were great and always willing to help me with anything I needed to fabricate. Finally, I'd like to thank Dr. Colsher and Dr. Chin for taking the time to serve on my thesis committee and provide valuable input. This would not have been possible without all of your help and accommodation.

1. Introduction

Imaging of ^{18}F-FDG with a PET/CT scanner has become a widely used diagnostic tool, especially for oncology patients (Ben-haim and Ell 2009). It has been used for many years to diagnose and stage many types of malignant cancer (Weber and Wieder 2006). Recently, there has been an increased focus on using quantitative measurements, such as the standard uptake value (SUV), of a tumor to assess a tumor's response to therapy. SUV is a semiquantitative measure of radiotracer accumulation in a region of the body, and is calculated by:

$$\mathrm{SUV} = \frac{\text{activity concentration in ROI}}{\text{injected activity/body size}} \tag{1.1}$$

The body size term can be represented by a few different quantities. In its simplest and most used form, it is the mass of the patient. Less commonly, it can refer to measures such as the surface area of the patient or lean body mass. SUV has been shown to be a more accurate metric of the tumor's response to therapy than tracking the changes in anatomic size of the tumor (W. A. Weber 2009).

A wide variety of factors influence SUV measurements. These factors can be broadly divided into two groups: biological factors and technical factors. While biological factors can greatly impact SUV measurements, the focus of this work is to minimize variation in small lesion quantitative measurements across different PET imaging systems. Due to this focus, only technical factors affecting SUV and quantitative measurements will be examined (Adams, et al. 2010).

An important fundamental technical factor is the intrinsic spatial resolution of the PET system. Differences in intrinsic spatial resolution arise due in part to differences in detector construction and ring size. This gives rise to partial volume effects, which affect the measured SUV of small objects by causing small objects to appear as larger, less intense objects. Large objects are less affected by this phenomenon (Shankar, Hoffman and Bacharach 2006).

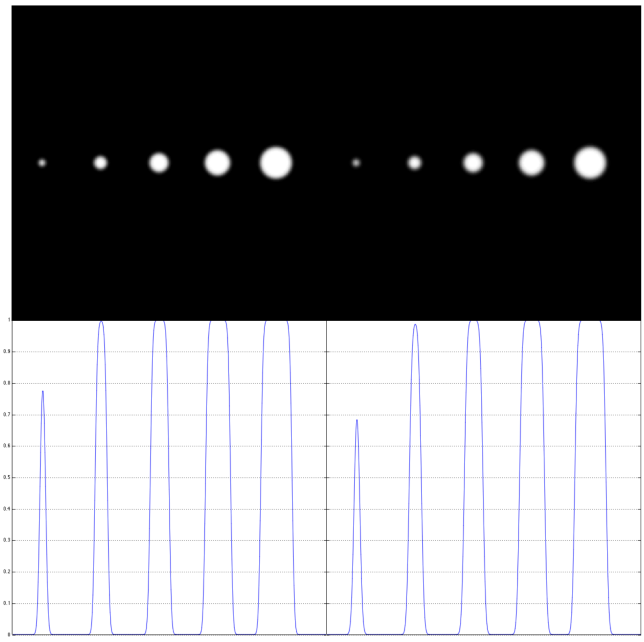

Figure 1: Simulated 1.0, 1.5, 2.0, 3.0, 4.0, and 5.0 cm spheres. The left image represents 7 mm of 2D blurring; the right represents 8 mm of 2D blurring.

There have been a few relatively recent developments that effectively improve spatial resolution. The first is modeling the point spread function (PSF) response of the PET system (Tong, Alessio and Kinahan 2010). This technique still has not been widely adopted, as there are still issues with effects such as Gibbs ringing. A second

development that has seen more widespread adoption is time-of-flight (TOF) imaging. Both of these have the potential to improve small lesion quantitation. However, their effect on quantitative measurements must be taken into account when comparing images produced by different systems. Even scanners of the same model have shown differences of up to 6% when imaging large uniform phantoms, due to variation in the radioactivity calibration (Scheuermann, et al. 2009).

Time-of-flight images tend to make images produced by TOF be quantitatively different from non-TOF images. The signal-to-noise ratio (SNR) improvement seen in TOF has been shown to be approximately dependent on the ratio of the size of the object being imaged (D) to the TOF spatial localization along the line-of-response (LOR), as shown in the equation below (Budinger 1983).

$$\text{TOF Gain} = \frac{D}{\Delta x} \tag{1.2}$$

This suggests that TOF imaging yields a larger improvement in large patients than small patients. One study experimentally investigated this theoretical improvement with a variable body size phantom. Although the main goal of this study was to investigate the effects of body size on SNR, it incidentally provides a look at the effects of body size on contrast. It showed that not only did TOF contrast measurements exhibit body size dependence, but also that conventional non-TOF measurements demonstrated a similar dependence on body size. The study additionally showed that recovered contrast decreased as body size increased for both TOF and non-TOF images. Finally, it

demonstrated that TOF recovered more contrast relative to non-TOF as body size increased (Wilson and Turkington, TOF-PET Small-Lesion Image Quality Measured Over a Range of Phantom Sizes 2013). A second factor unique to TOF imaging is better cold region recovery compared to hot regions. One study has shown that TOF has approximately 2 times better recovery of hot regions and 4 times better recovery of cold regions than non-TOF in a medium to large sized abdomen sized phantom (Smith 2011).

Image reconstruction parameters will also affect measured quantitation. These include: image matrix size, field of view (FOV) size, number of iterations, and post-smoothing. Matrix size and FOV size both affect the physical size the voxels represent. A larger matrix decreases the size of voxels. Studies have shown this tends to increase SUV measurements in small lesions at the expense of more uncertainty in any single voxel. A smaller FOV also effectively decreases the physical size a voxel represents (Adams, et al. 2010) (Westerterp, et al. 2007).

Increasing the number of iterations that the iterative reconstruction algorithm performs increases spatial resolution at the expense of increased image noise. TOF creates less noisy images for the same resolution relative to non-TOF images. Past a certain number of iterations (dependent on the specific imaging context) the resolution plateaus while image noise continues to increase (Adams, et al. 2010) (Jaskowiak, et al. 2005).

SUV can also be affected by the placement of the lesion. A smaller SUV will be measured if a lesion is not centered in a voxel, so if care is not used in precise alignment of the phantom then one machine could measure a higher SUV_{max} relative to another due to better centering of the lesions in the voxels (Adams, et al. 2010). This effect is not controllable in patients.

Object size is very important to consider when attempting to quantify the object. Adams, et al. showed that the effect these reconstruction parameters had on 2.5 cm spheres was greatly diminished compared to 1.0 cm spheres. However, they concluded that small 1.0 cm lesions were still clinically relevant for early detection of cancer as well as for large, diffuse tumors containing large amounts of necrotic tissue interspersed with small regions of active cancer cells. The latter case is especially important now that physicians are using SUV to judge a cancer's response to treatment (Adams, et al. 2010).

All of these unaccounted-for factors show a significant need for methods to ensure quantitative lesion measurements are not dramatically different from one scanner to another. Due to the difficulty of accounting for all of these factors, our nuclear medicine department at Duke currently scans each patient on the same PET system they were originally scanned on, if possible. Additionally, these factors make multi-clinic studies involving quantitative PET measurements difficult. Minimizing the differences in lesion measurements between PET systems would allow for greater flexibility in

choosing which system to scan patients on and increase the viability of large multi-center studies involving quantitative measurements.

Efforts to standardize quantitative measurements across PET systems have been studied by Makris, et al. They showed it was possible to harmonize SUV and recovery coefficient (RC) measurements across scanners using commercially available phantoms (Makris, et al. 2013). However, their study was with just one system. Varying the number of iterations simulated multiple systems. So far, no study on the viability of standardization of small object quantitation has been published, where full recovery of object's activity concentration is impossible due to partial volume effects.

One potential challenge to success was that the phantoms used for calibration had axial symmetry, insensitive to axial resolution. Quantitative measurements of spheres in the test phantom, and of any realistic small lesions, are sensitive to axial resolution. This work examines a best-case scenario where the axial resolutions of the two systems are similar. Systems with larger axial resolution differences would likely need to have their axial resolutions cross-calibrated first with a phantom such as the Defrise disk phantom. A second potential challenge to success is the body size dependence of PET for quantitative measurements.

The aim of this work was to establish and validate a method using commercially available and easily filled and scanned phantoms to cross-calibrate the image processing protocols for two different systems with the goal that small lesion measurements can be

standardized across PET systems. For this work, a five percent difference in simulated lesion measurements after cross-calibration was considered acceptable, and a two percent difference was considered equivalent.

2. Methods and Materials

The basic procedure performed to cross-calibrate the systems was to scan a Jaszczak phantom (Data Spectrum Corporation 2007) on both systems, reconstruct the images with a range of iterations or post-smoothing, then manipulate reconstruction parameters to equalize the Jaszczak images. These cross-calibrated parameters were then assessed for their performance in equalizing small lesion quantitation using quantitative measurements from phantoms filled with simulated lesions and reconstructed with the cross-calibrated parameters. Two different methods of cross-calibrating were evaluated: cross-calibration via iteration control and cross-calibration via post-smoothing. This section will detail, in order, the PET imaging systems used, the phantoms used, image reconstruction parameters, contrast measures, and finally the different cross-calibration methods.

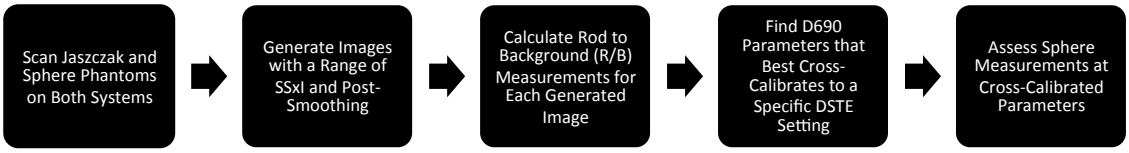

Figure 2: A basic flow chart of the experimental method.

2.1 PET Systems

Two PET/CT systems were used in this experiment: a GE Discovery 690 (D690) and a GE Discovery STE (DSTE). The systems were located in the same radiology suite, making it possible to do scans of the same phantom on both machines successively

without radiotracer decay substantially affecting the results. Overall, the two systems possess similar physical performance. Both systems have the same axial detector crystal dimension, resulting in very similar spatial resolution. The main difference between the two is the D690 detector crystals are slightly smaller in the transaxial dimension, resulting in slightly better transaxial spatial resolution for the D690. In addition, the D690 has TOF capabilities.

2.2 Abdomen Phantom filled with Simulated Lesions

A custom-built abdomen phantom filled with ten 1.0 cm hot spheres to simulate small lesions served as for the basis for comparison. Although this phantom does a very good job simulating small lesions, it would not be suitable for studies involving multiple sites due to the difficult and time-consuming process of filling the phantom. Additionally, it would be difficult to ensure precise, repeatable sphere to background activity concentration ratios. The ability to fill the phantom once and then scan on both systems negated these issues for this study.

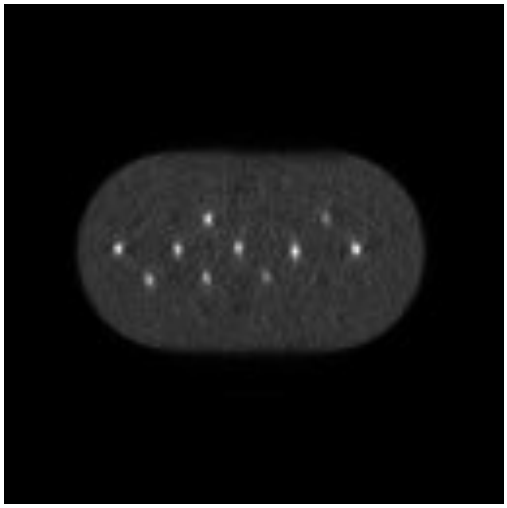

Figure 3: A PET image of the abdomen phantom.

The spheres were filled with a ^{18}F-FDG solution 8 times greater than the warm background, along with a iodine-based CT contrast agent to facilitate an automatic region of interest (ROI) analysis of the spheres from the CT images. Approximately 2 mCi of radiotracer was in the phantom at the time of the scans. A 20-minute PET scan was performed sequentially on both systems. The scans were 20 minutes long in order to minimize image noise.

2.3 Jaszczak Phantom

Jaszczak inserts are made of 6 wedge shaped sections, with each section being composed of different uniformly sized plastic rods. This is known as a "cold rod" phantom because the non-radioactive rods create areas with no activity. A second type of Jaszczak insert consists of a cylindrical plastic block with holes drilled wherever a plastic rod existed in the cold rod phantom, creating what is known as a "hot rod"

phantom. These phantoms can be seen in Figure 4. The phantoms were filled with approximately 1 mCi of ^{18}F-FDG and scanned for 20 minutes sequentially on both systems.

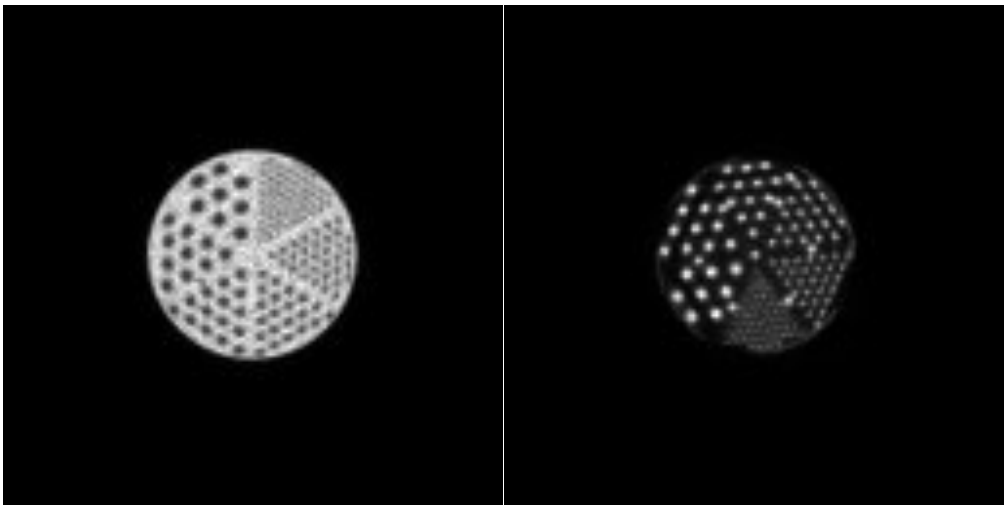

Figure 4: PET images of the cold rod Jaszczak phantom (left), and the hot rod phantom (right).

2.4 Jaszczak Phantom with Hot Spheres

Due to dissatisfaction with preliminary results in cross-calibrating non-TOF images to TOF images, a custom built insert for the Jaszczak phantom was made to evaluate the effects of the object size dependence of PET on the predictive power of the Jaszczak phantoms. The custom built device contained mounting points for six hot spheres used in the abdomen phantom and was inserted into the Jaszczak phantom, which also contained the hot rod insert. The spheres were filled with an 8 to 1 activity concentration ratio relative to background and CT contrast, as was done in the abdomen

phantom. The phantom was filled with approximately 1 mCi of ^{18}F-FDG, then scanned for 20 minutes on each machine.

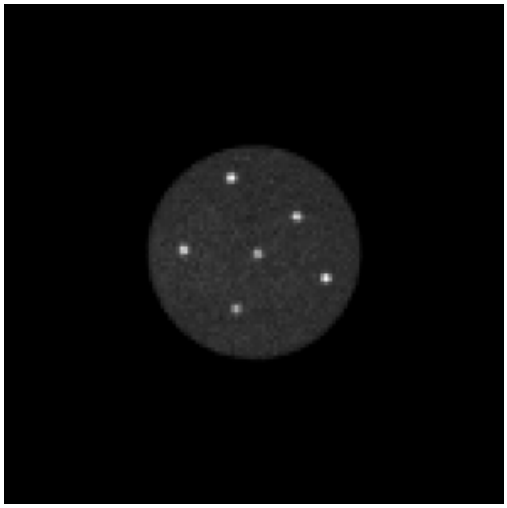

Figure 5: A PET image of the hot spheres inside the Jaszczak phantom.

2.5 Abdomen Phantom with Jaszczak Insert

Finally, as another way to attempt to evaluate TOF to non-TOF correlation, a hot rod Jaszczak insert was attached to one of the endplates of the abdomen phantom. The rest of the phantom was filled as outlined in the abdomen phantom section, and scanned for 20 minutes.

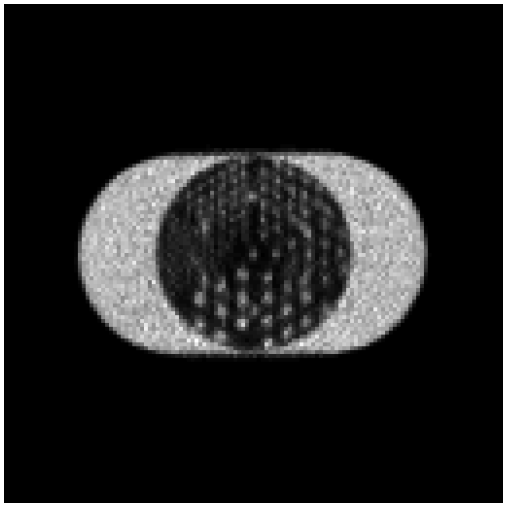

Figure 6: A PET image of the hot rod Jaszczak insert inside the abdomen phantom.

2.6 Reconstruction Parameters

The images were reconstructed while making every reconstruction parameter as similar as possible. This included matrix size (128 by 128), random event correction method (by singles), attenuation correction (CT-based), scatter correction (model-based), dead time correction (on), and iterative reconstruction method (OS-EM). We will refer to subsets and iterations as IxSS, as the product of OS-EM subsets and iterations is comparable to the total number of ML-EM iterations for lower numbers of iterations (Hudson and Larkin 1994). No transaxial post-smoothing was done initially, and both scanners used a 1-4-1 axial filter kernel across adjacent slides. The images were reconstructed with a range of subsets and iterations, shown in Table 1. Additionally, two different sets of IxSS were chosen to post-smooth with a range of smoothing. This is detailed more in section 2.8.2. Finally, the sphere scans were replayed as 1, 2, 3, 4, and 5

minute scans to evaluate how higher noise situations affect the performance of the cross-calibrated parameters. The start times for the replays were made as similar as possible to minimize the effect of tracer decay on the results.

Table 1: IxSS parameters used to reconstruct the images in this work.

D690 (non-TOF and TOF)	DSTE
3x1	3x1
3x2	3x2
3x3	3x4
3x4	3x5
3x6	3x7
3x8	3x8
3x9	3x10
3x12	3x14
3x16	3x20
3x18	3x28
3x24	3x35
3x32	3x40
3x36	3x56
3x48	3x70

2.7 Contrast Measurement Methods

2.7.1 Sphere Contrast

A CT based analysis of the hot spheres was performed, as detailed in J M Wilson's paper, "Multisphere phantom and analysis algorithm for PET image quality assessment" (Wilson and Turkington, Multisphere phantom and analysis algorithm for PET image quality assessment 2008). The program finds each sphere in the CT image and creates a volume of interest (VOI) in the PET image equal to the dimensions of the

sphere at the location found in the CT image. This program records the average maximum value and average mean value across all the spheres, along with the standard deviation of both measurements. These values were then normalized relative to background ROIs to correct for decay. Strictly speaking these measurements are not SUV, as they use a different normalization factor. Normalizing the phantom background, as was done in this study, minimizes any effect of differences in absolute calibration accuracy of the different scanners.

2.7.1 Jaszczak Phantom Contrast

For this study, a semi-automatic CT-based analysis of images from the Jaszczak phantoms was developed. From the CT scan, a binary mask of rod locations was created. The value 1 represented the presence of a rod and 0 represented the absence. For the cold rod phantom, this meant thresholding the image for plastic (90 to 150 HU), and for the hot rod phantom this meant thresholding for water (-15 to 25 HU). The PET image was then scaled up to the size of the CT image, using a nearest neighbor method of up-scaling. The up-scaled PET image was then multiplied by the rod mask, resulting in a PET image of just the rods. From these masked images the average activity concentration of the rods was calculated. These values were normalized relative to a background ROI, creating a rod to background ratio (R/B) measurement. This rod to background ratio served as a potential small lesion contrast (and therefore, resolution) surrogate.

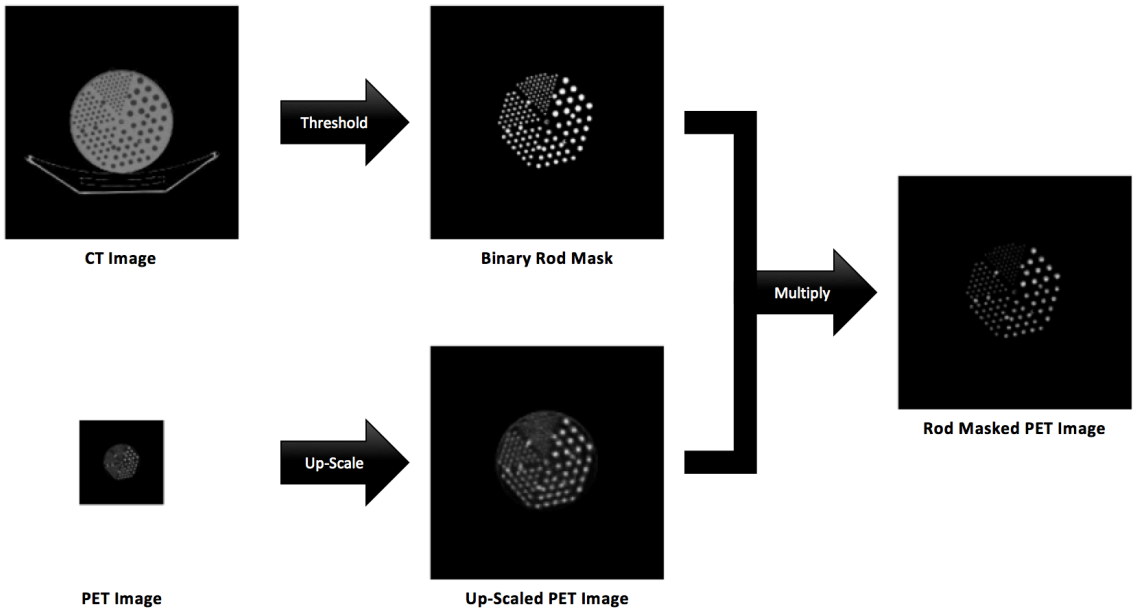

Figure 7: Overview of the CT-based Jaszczak phantom contrast analysis.

2.8 Cross-Calibration Methods

2.8.1 Iteration Control

Multiple tests were performed to evaluate the performance of the Jaszczak R/B measurement as a lesion contrast surrogate. The first test was to calculate the number of iterations needed on the D690 to provide as equal as possible R/B measurements for a variety of iterations on the DSTE, and then evaluate the difference in sphere measurements using these iteration levels. To do so, the images were reconstructed at a variety of iterations, and then linear interpolation was used to find the number of iterations that best matched the D690 Jaszczak images for a given number of iterations of the DSTE images.

2.8.2 Post-Smoothing

The second test was to evaluate the performance of this R/B ratio as a lesion contrast surrogate when using post-smoothing (PS) instead of iterations to control contrast. To do so, unsmoothed images were reconstructed with two levels of iterations. The first was similar to those used clinically on these systems for SBMT scans. The DSTE images used 42 iterations, the non-TOF D690 48 iterations, and the TOF D690 36 iterations. The second was at high iterations where contrast has started to plateau. For the D690 this was 96 iterations for both TOF and non-TOF, and 105 iterations for the DSTE. Images of the hotrod Jaszczak and the abdomen phantom produced by these reconstruction parameters were successively smoothed in increments of 0.5 mm from 0 mm to 10 mm of PS. These smoothed Jaszczak images were analyzed by the semi-automatic CT-based analysis outlined at the beginning of this section to obtain the R/B ratios at each smoothing level. Linear interpolation was used to calculate the amount of smoothing needed to match the R/B ratios of the D690 Jaszczak image sets to the DSTE Jaszczak image sets for each level of DSTE image smoothing. These smoothing levels were then applied to sphere images, and the resulting quantitative measurements of the spheres were compared.

Finally, the ability of PS to further minimize the differences in sphere measurements after coarsely minimizing the differences in sphere measurements with iteration control was evaluated. This was evaluated by successively smoothing similar

image sets obtained by iteration control and observing the change in the difference of sphere measurements.

3. Results

3.1 Abdomen Phantom Measurements with No Cross-Calibration

Figures 8 and 9 below show the average mean value and average maximum value of the 1 cm spheres for a range of iterations. The D690 tended to have a higher recovery of activity than the DSTE for a given number of iterations. The D690 TOF images exhibited a much higher recovery of activity than both the non-TOF D690 images and the DSTE images.

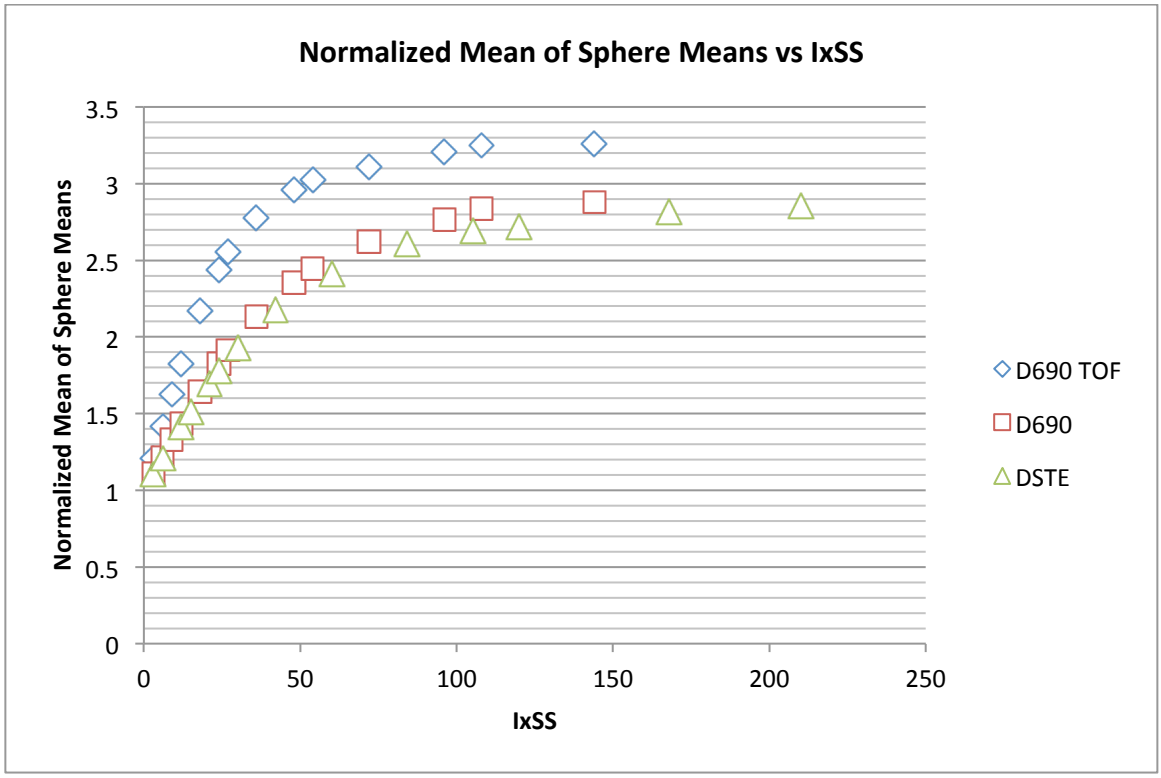

Figure 8: The normalized mean of sphere means as a function of iterations for the two systems.

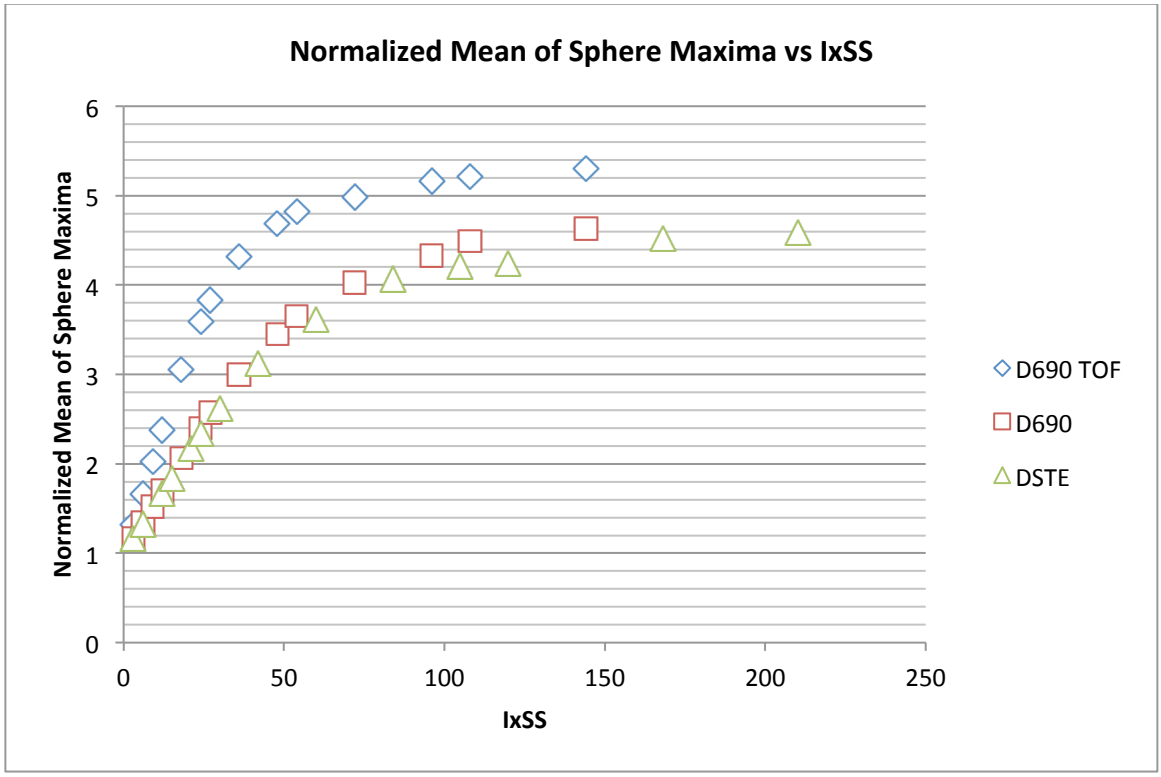

Figure 9: The normalized mean of sphere maxima as a function of iterations for the two systems.

3.2 Cross-Calibration via Iteration Control

3.2.1 Jaszczak Phantom Correlation with Spheres Inside the Abdomen Phantom

The hot rod phantom somewhat outperformed the cold rod phantom as the basis for cross-calibrating the system reconstruction parameters. In the range clinically used for SBMT and leg scans at Duke (30-42 iterations), the absolute sphere mean difference was less than 1% when using the hot rod predicted iterations. The cold rod phantom cross-calibration over this range resulted in a 1.6% difference, still quite acceptably good.

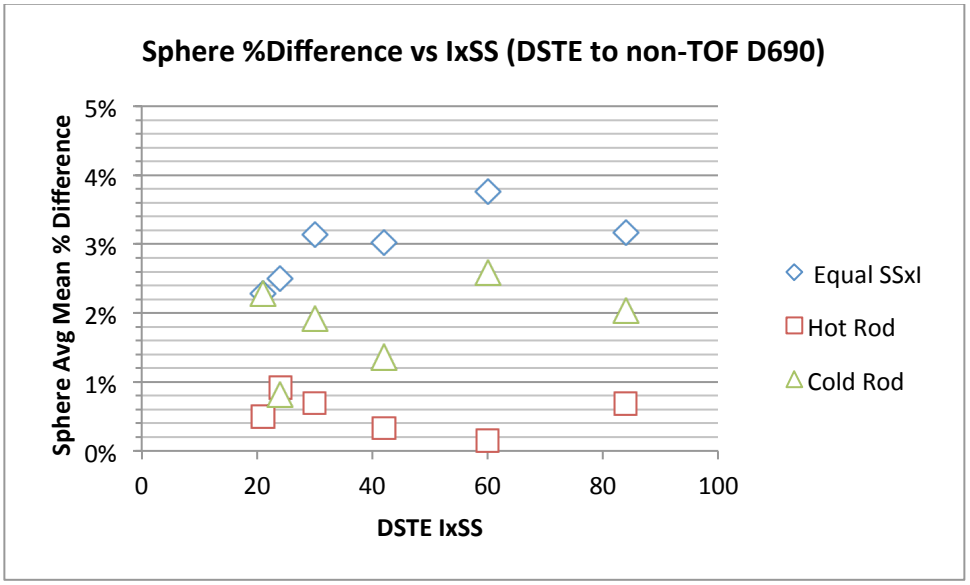

Figure 10: Assessment with spheres inside the abdomen phantom of the Jaszczak phantoms' cross-calibration performance via iteration control between the DSTE and non-TOF D690 systems.

Figure 11 shows neither phantom did very well cross-calibrating the number of iterations needed for the TOF D690 sphere quantitation to match the DSTE sphere quantitation when using the spheres in the abdomen phantom. Both performed better than setting iterations equal, with the hot rods resulting in a 6% average sphere mean difference and the cold rods resulting in a 11% average sphere mean difference.

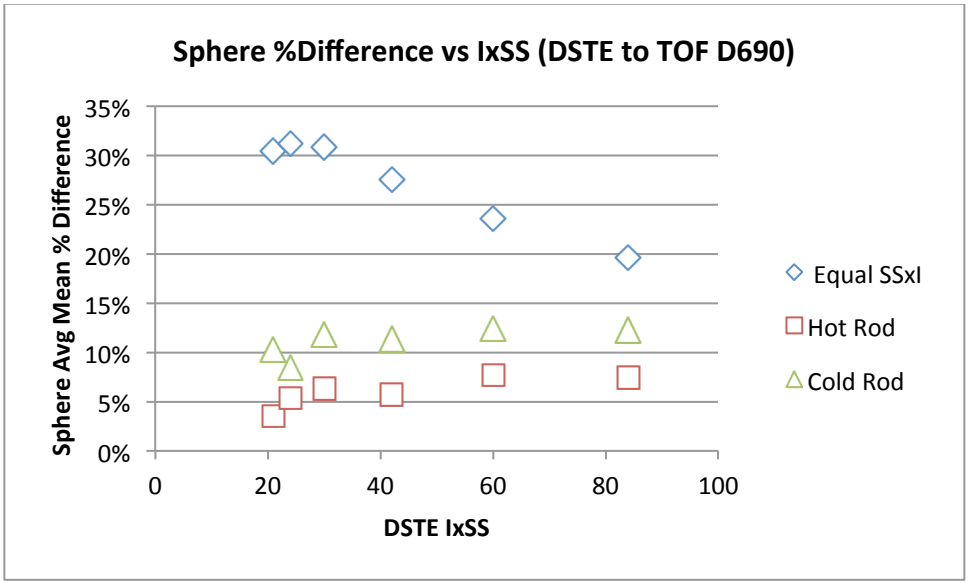

Figure 11: Assessment with spheres inside the abdomen phantom of the Jaszczak phantoms' cross-calibration performance via iteration control between the DSTE and TOF D690 systems.

3.2.2 Jaszczak Phantom Correlation with Spheres Inside the Jaszczak Phantom

For non-TOF image sets, both Jaszczak inserts cross-calibrated the number of iterations needed to minimize the difference in sphere measurements very well when comparing to the spheres placed inside the Jaszczak phantom. Figure 12 shows sphere measurements were actually slightly more different on average when using the iterations predicted by both Jaszczak inserts than when setting the number of iterations equal. However the average difference was still less than 3%.

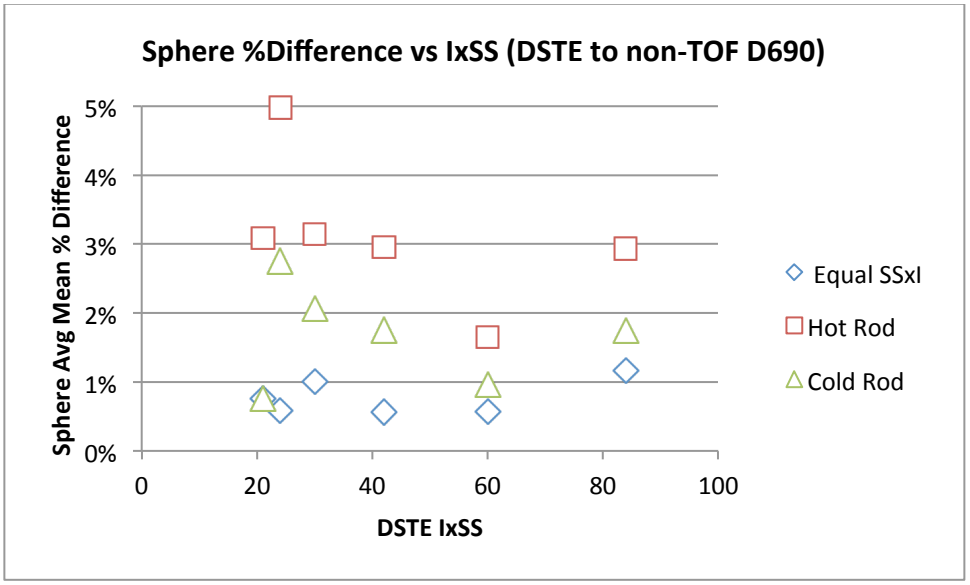

Figure 12: Assessment with spheres inside the Jaszczak phantom of the Jaszczak phantoms' cross-calibration performance via iteration control between the DSTE and non-TOF D690 systems.

Using the hot rod phantom cross-calibration for TOF D690 and DSTE image sets resulted in a mere 3.5% average difference in sphere means at iterations greater than 10. The cold rod phantom provided slightly better results at very low iterations, and actually performed slightly better on average than the hot rod phantom.

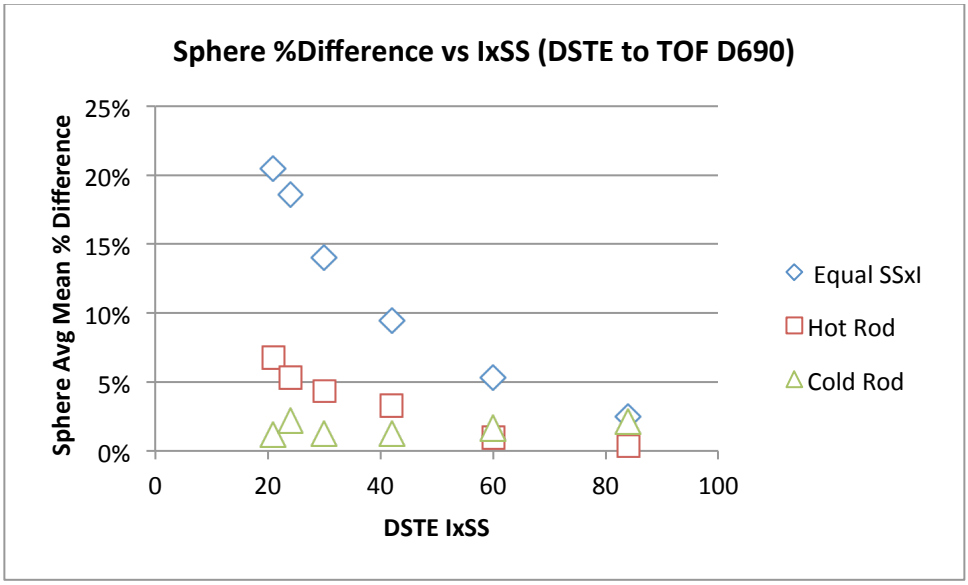

Figure 13: Assessment with spheres inside the Jaszczak phantom of the Jaszczak phantoms' cross-calibration performance via iteration control between the DSTE and TOF D690 systems.

3.2.3 Jaszczak Phantom Correlation when Placed Inside the Abdomen Phantom

The Jaszczak phantom placed inside the abdomen phantom provided unusable data. One reason for this failure is due to the activity surrounding the Jaszczak insert affecting the insert itself. When the Jaszczak phantom is used in its normal container, practically no background activity surrounds the circumference of the insert. In the abdomen phantom, there was a very large region of background encircling the Jaszczak insert. This caused poorer resolution to actually have higher R/B ratios up to a certain number of iterations, since more of the surrounding activity was blurred into the rods. This effect can be seen in Figure 14.

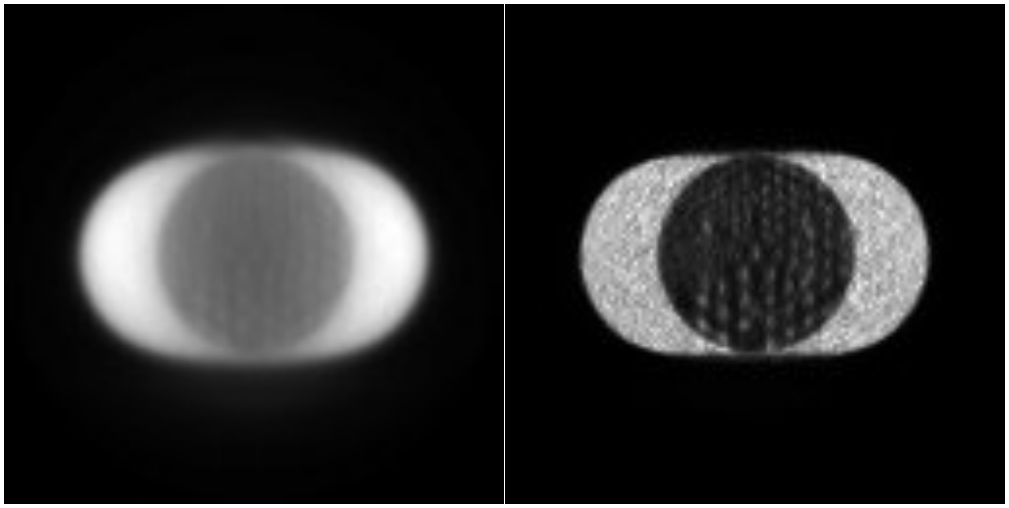

Figure 14: Images of the Jaszczak insert inside the abdomen phantom. The left image is a low IxSS image. The right image is a higher IxSS image, similar to what is used clinically. Blurring of the background activity into the rods can be plainly seen in the left image.

3.3 Cross-Calibration via Post-Smoothing

3.3.1 Jaszczak Phantom Correlation with Spheres Inside the Abdomen Phantom

3.3.1.1 Clinical Iterations

As Figure 15 shows, the cross-calibrated smoothing parameters from both Jaszczak phantoms consistently underestimated the amount of smoothing required for the sphere measurements from the D690 abdomen phantom images to match measurements from the DSTE images at a given smoothing level. However, their cross-calibrated smoothing levels resulted in much more similar measurements than applying equal smoothing.

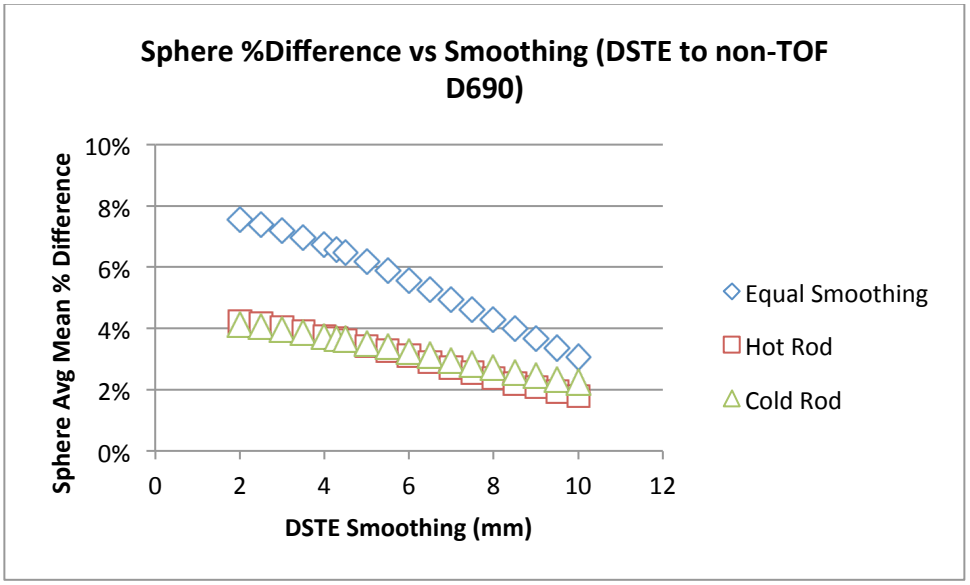

Figure 15: Assessment with spheres inside the abdomen phantom of the Jaszczak phantoms' cross-calibration performance via smoothing control at clinical iterations between the DSTE and non-TOF D690 systems.

Figure 16 shows that neither phantom performed well at cross-calibrating the amount of smoothing the TOF D690 abdomen images needed to match the sphere quantitation of the DSTE abdomen images. Their overall performance was similar and much better than equal smoothing, but the difference in sphere measurements was still about 12.5% different.

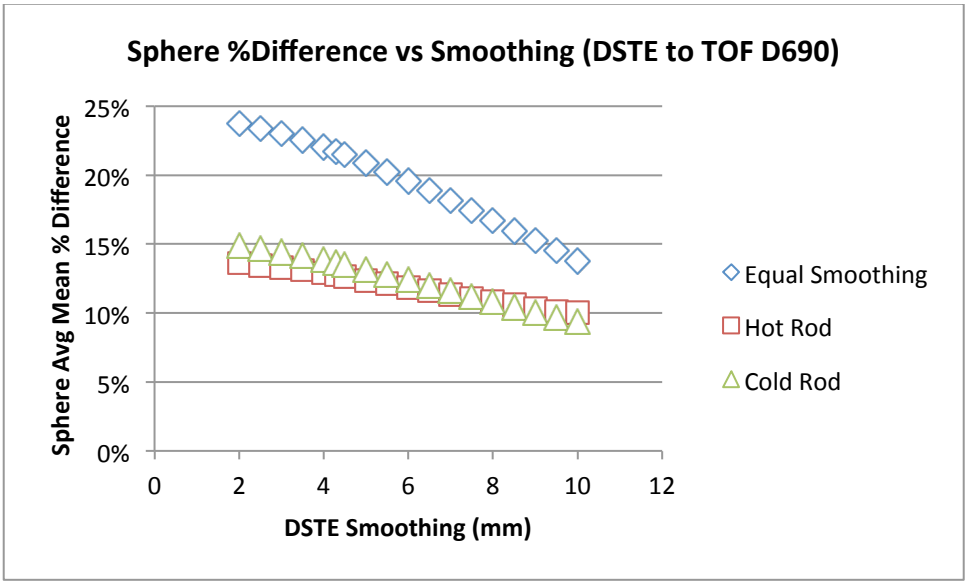

Figure 16: Assessment with spheres inside the abdomen phantom of the Jaszczak phantoms' cross-calibration performance via smoothing control at clinical iterations between the DSTE and TOF D690 systems.

3.3.1.1 High Iterations

The cross-calibration from both Jaszczak phantoms provided smaller differences in sphere quantitation at high iterations than at clinical iterations, as Figure 17 illustrates. Both performed slightly worse than setting equal smoothing, but the average difference in sphere measurements was still less than 3%.

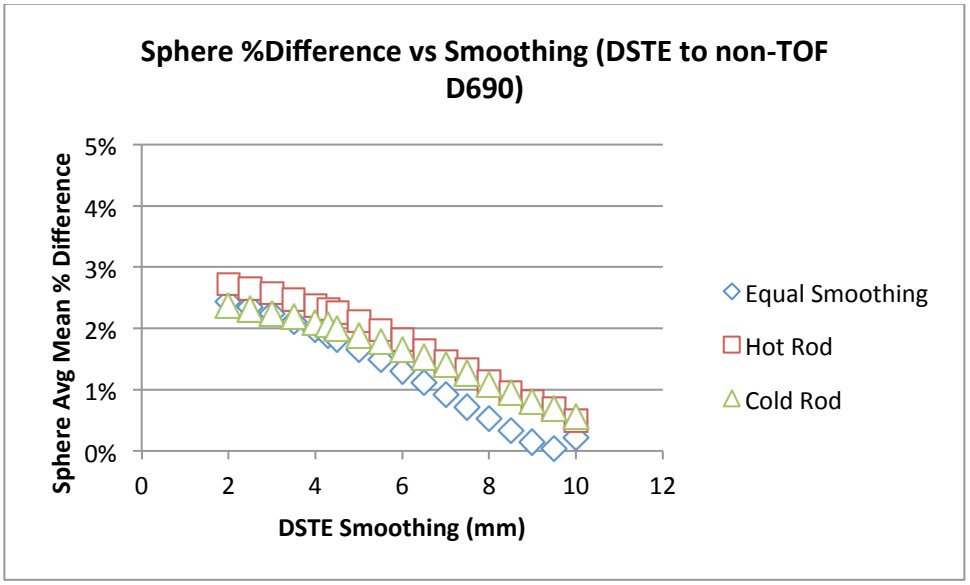

Figure 17: Assessment with spheres inside the abdomen phantom of the Jaszczak phantoms' cross-calibration performance via smoothing control at high iterations between the DSTE and non-TOF D690 systems.

Similarly, the TOF D690 and DSTE images were much more quantitatively similar when using the Jaszczak cross-calibrated smoothing levels at high iterations than clinical iterations. Both brought the average difference in sphere quantitation down from approximately 13% at equal smoothing to approximately 8.5%. This is still much worse than the non-TOF D690 to DSTE quantitative differences when using the Jaszczak cross-calibrated smoothing amounts.

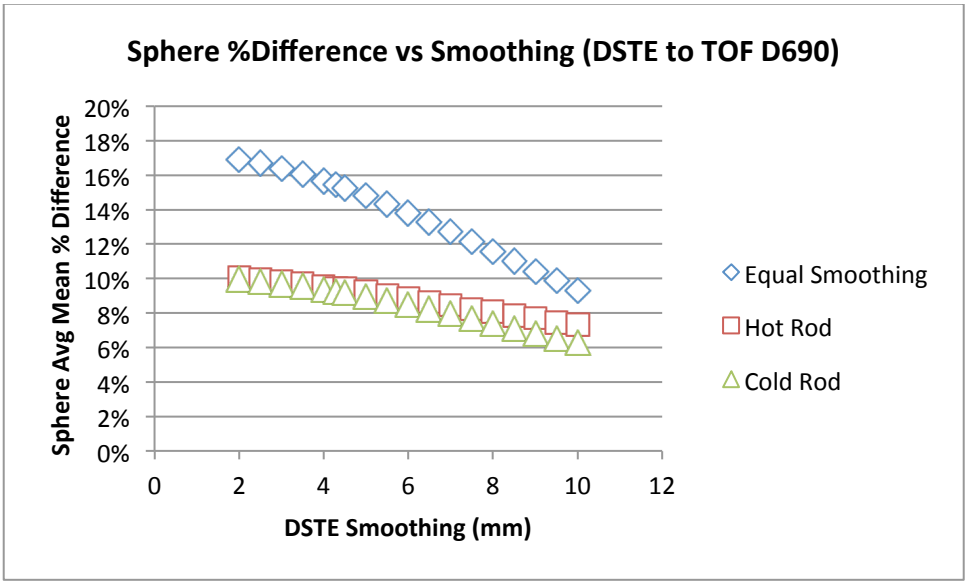

Figure 18: Assessment with spheres inside the abdomen phantom of the Jaszczak phantoms' cross-calibration performance via smoothing control at high iterations between the DSTE and TOF D690 systems.

3.3.2 Jaszczak Phantom Correlation with Spheres Inside the Jaszczak Phantom

3.3.2.1 Clinical Iterations

Figure 19 shows that cross-calibration of both inserts correlated much better with the spheres placed inside the Jaszczak phantom than the spheres inside the abdomen phantom. Both inserts performed better than equal smoothing, with an average sphere measurement difference of less than 2%.

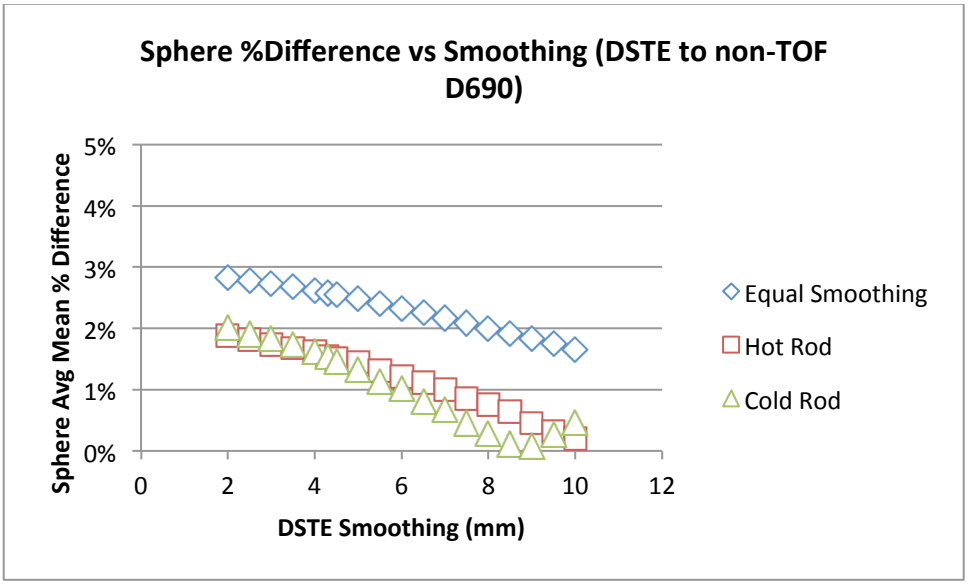

Figure 19: Assessment with spheres inside the Jaszczak phantom of the Jaszczak phantoms' cross-calibration performance via smoothing control at clinical iterations between the DSTE and non-TOF D690 systems.

Similarly, Figure 20 shows that Jaszczak inserts performed quite well at cross-calibrating the amount of smoothing needed to equalize sphere measurements between TOF and non-TOF images. The cross-calibrated smoothing cut the average sphere difference from approximately 6.5% at equal smoothing to approximately 3% using the predicted smoothing amounts.

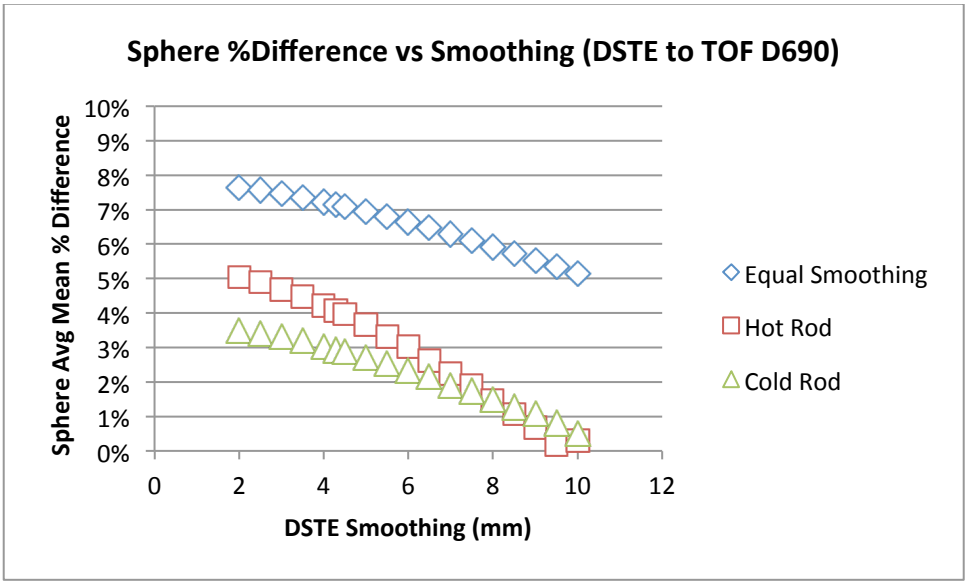

Figure 20: Assessment with spheres inside the Jaszczak phantom of the Jaszczak phantoms' cross-calibration performance via smoothing control at clinical iterations between the DSTE and TOF D690 systems.

3.3.2.2 High Iterations

The Jaszczak phantom cross-calibrations correlated much better with spheres placed inside of the Jaszczak phantom container, thereby equalizing body size. Figure 21 shows that although simply equalizing smoothing provided less different sphere measurements, using the cross-calibrated smoothing amounts still provided sphere measurements that were less than 2% different on average.

The cold rod Jaszczak phantom performance was similar to that of the hot rod phantom for high iterations. Like the hot rod phantom, simply performing equal smoothing with the DSTE and non-TOF D690 image sets resulted in more similar sphere measurements than using the smoothing amounts predicted by the cold rod Jaszczak phantom.

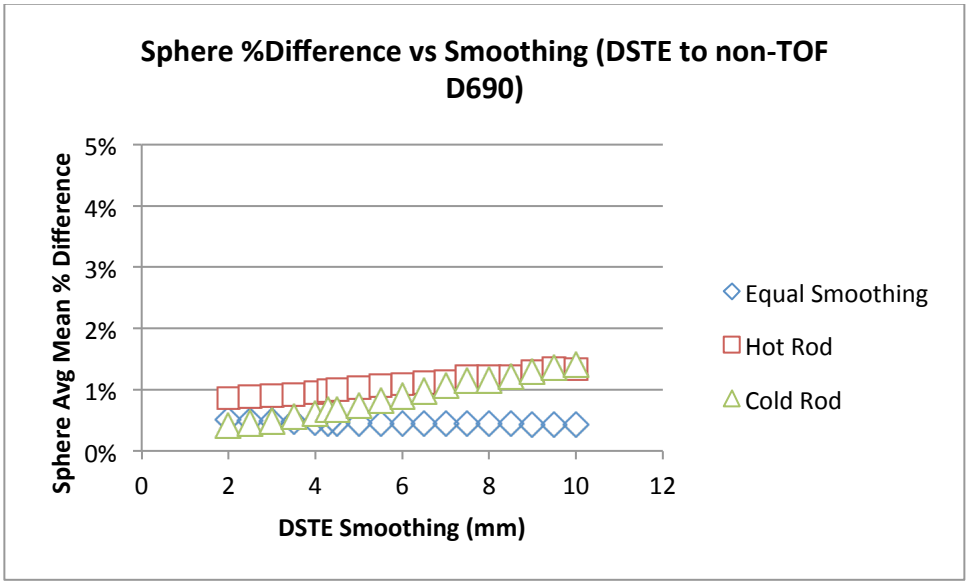

Figure 21: Assessment with spheres inside the Jaszczak phantom of the Jaszczak phantoms' cross-calibration performance via smoothing control at high iterations between the DSTE and non-TOF D690 systems.

Figure 22 shows that using the Jaszczak cross-calibrated smoothing is much more beneficial when the starting data points are more disparate. In this case, the DSTE and TOF D690 image sets were compared, providing a larger initial difference in resolution. Using the cross-calibrated smoothing resulted in reducing the difference in sphere measurements by more than half. The average difference with equal smoothing resulted in an average sphere measurement difference of 3.97%. With the predicted amount of smoothing, this average difference was reduced to 1.78%.

The DSTE to TOF D690 performance of the cold rod phantom was similar to the hot rod phantom. Using the predicted amount of smoothing resulted in a halving of the difference in sphere measurements when compared to using equal smoothing.

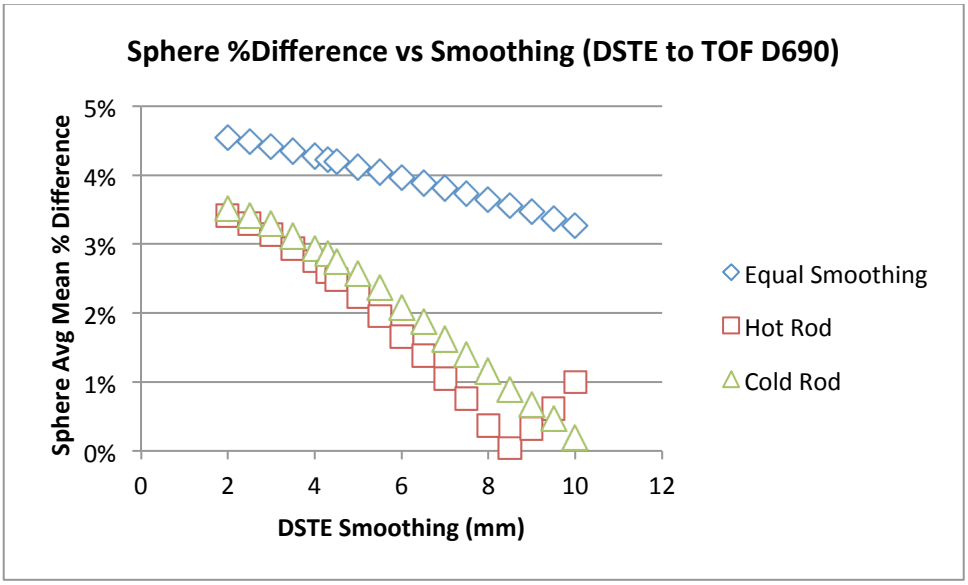

Figure 22: Assessment with spheres inside the Jaszczak phantom of the Jaszczak phantoms' cross-calibration performance via smoothing control at high iterations between the DSTE and TOF D690 systems.

3.3.3 Identical Post-Smoothing after Cross-Calibrating Iterations

To evaluate how well post-smoothing minimizes the differences between similar images, a pair of iterations resulting in similar sphere measurements between the D690 and DSTE were chosen. The D690 image set was 54 iterations, and the DSTE image set was 60 iterations. As the figure below shows, the D690 and DSTE mean sphere values start to converge at higher PS levels. The percent difference in the sphere means is almost halved when using smoothing levels near what is used clinically (6.4 mm).

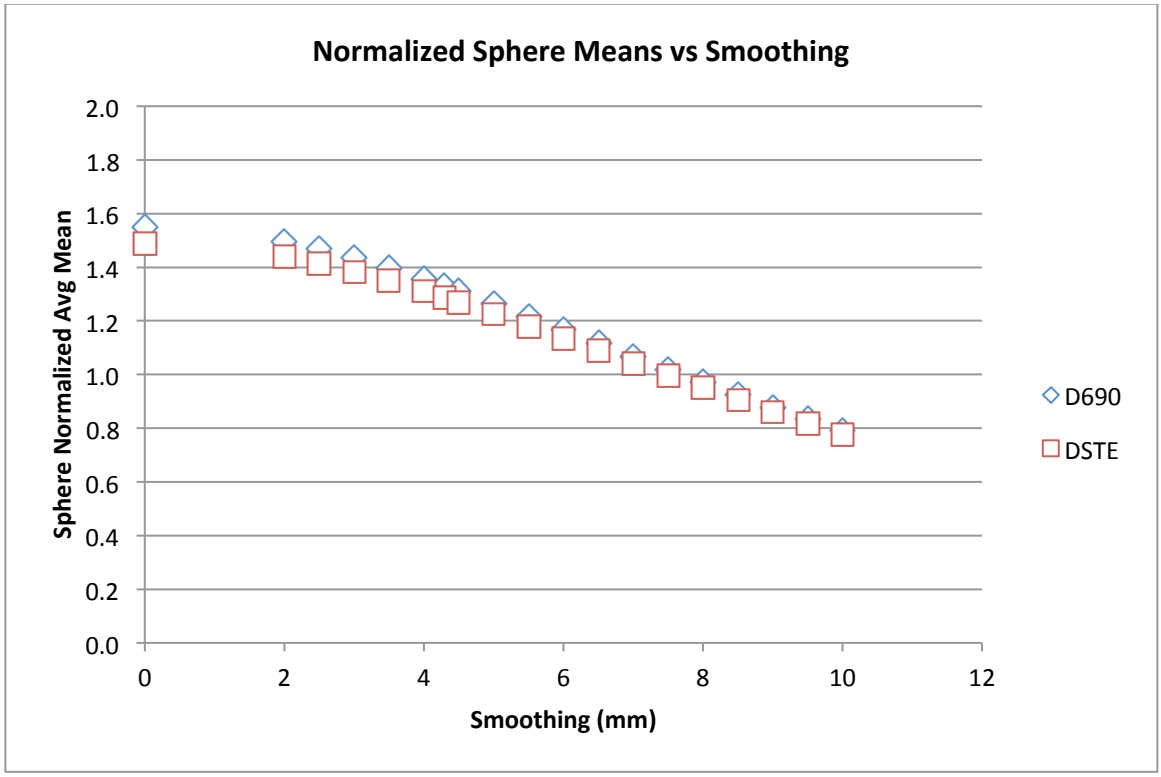

Figure 23: An illustration of the effects of post-smoothing on similar images.

3.4 Cross-Calibration evaluated with Noisy Images

As mentioned in section 2.6, the sphere scans were replayed at shorter scan durations to simulate noisier datasets, which are more representative of the noise levels seen clinically. Since both the hot and cold rod Jaszczak phantoms yielded similar results in earlier sections, only the cross-calibrations obtained from the hot rod Jaszczak phantom were examined in this section. Figure 24 shows that the shorter abdomen scan durations resulted in sphere differences of mostly similar magnitude when cross-calibrating with iteration control, however the sphere measurements from the shorter scan durations were much more different compared to the twenty-minute scan. The

one-minute abdomen scan was the only scan duration that behaved substantially differently from the other short scan durations, with the sphere differences becoming quite different at higher iterations. The spheres inside the Jaszczak phantom showed no substantial difference across scan durations.

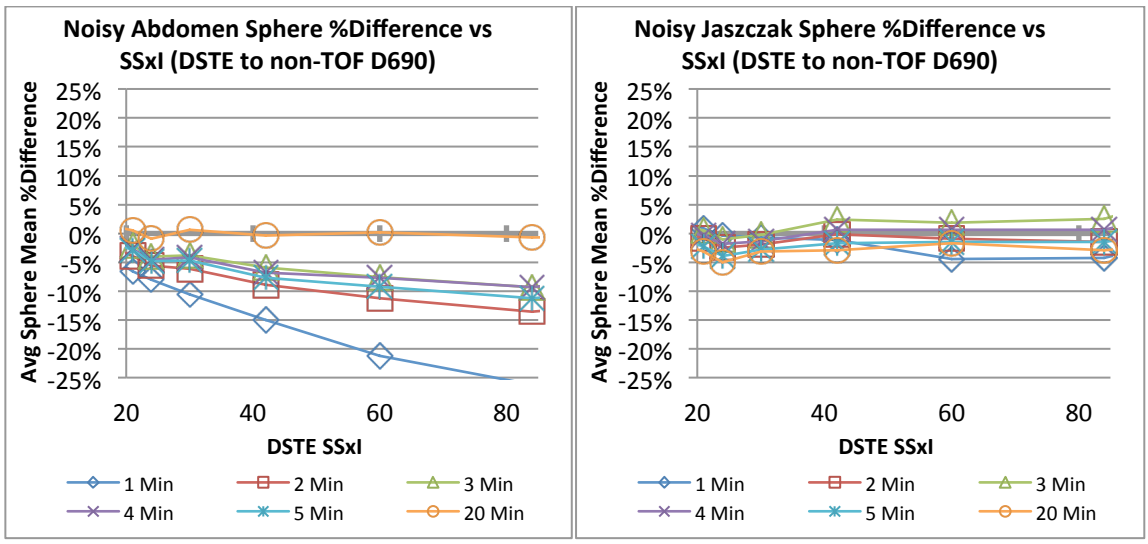

Figure 24: A comparison of the iteration control cross-calibrated sphere differences of shorter scan durations with the original 20-minute scans.

When cross-calibrating with post-smoothing, the one-minute abdomen scan duration once again behaves substantially differently from the other short abdomen scan durations at both clinical and high iterations. For the spheres inside the Jaszczak phantom there was a larger difference in sphere measurements of the shorter scan durations compared to the twenty-minute scan, with the sphere differences at high IxSS levels being unacceptably large.

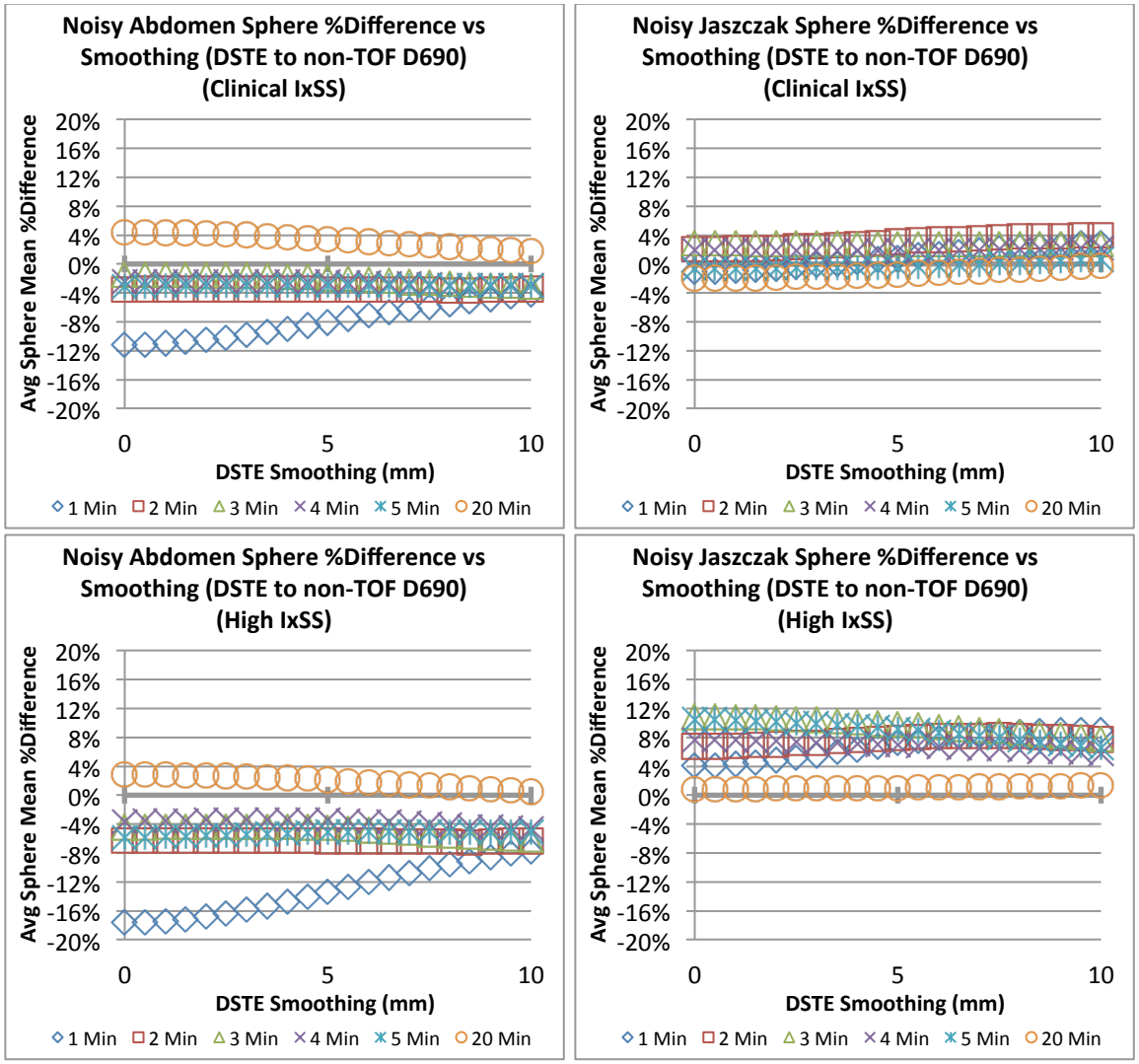

Figure 25: A comparison of the post-smoothing control cross-calibrated sphere differences of shorter scan durations with the original 20-minute scans.

The differences in sphere measurements behaved differently when cross-calibrating the DSTE with the TOF D690 images with iteration control, as shown in Figure 26. Now both the one and two minute abdomen scan durations behaved differently from the other short scan durations, which behaved very similarly to each other. All scan durations behaved similarly for the spheres inside the Jaszczak phantom.

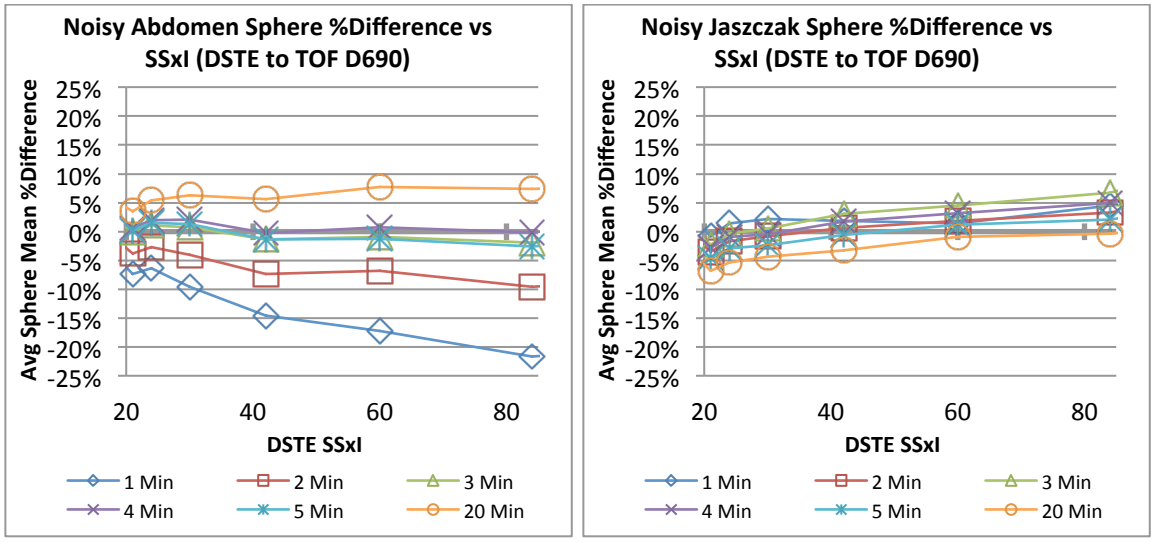

Figure 26: A comparison of the iteration control cross-calibrated sphere differences of shorter scan durations with the original 20-minute scans between the DSTE and TOF D690.

Sphere measurements of the post-smoothing cross-calibrated DSTE and TOF D690 images shown in Figure 27 behaved similarly to the post-smoothing cross-calibrated DSTE to non-TOF D690 images shown in Figure 25. The biggest change in behavior between the non-TOF and TOF cross-calibrations is a larger difference in sphere measurements in the abdomen phantom for the TOF cross-calibration. As seen in

the non-TOF to TOF iteration control cross-calibration, the one and two minute abdomen scans behaved differently from the other short duration scans.

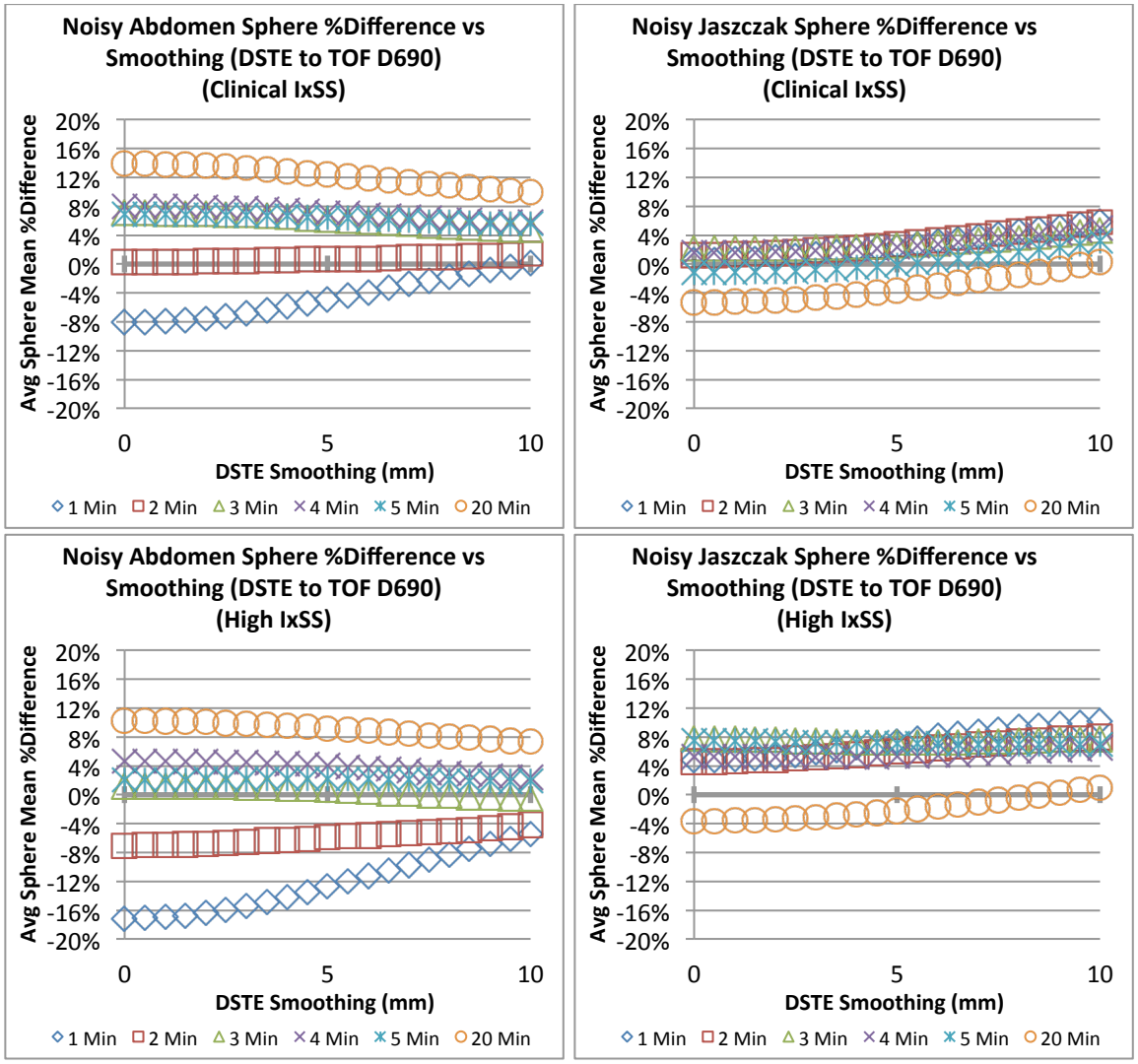

Figure 27: A comparison of the post-smoothing control cross-calibrated sphere differences of shorter scan durations with the original 20-minute scans between the DSTE and TOF D690.

4. Discussion

4.1 Abdomen Phantom Measurements with no Cross-Calibration

The results from the abdomen phantom data agreed with expectations. Figures 8 and 9 showed that the D690 images exhibited higher sphere contrast than the DSTE images for a given number of iterations, owing to its slightly higher axial resolution. Similarly, the TOF D690 images exhibited higher sphere contrast than the non-TOF D690 or DSTE images for a given number of iterations due to TOF effectively increasing resolution. This shows that simply setting IxSS or PS equal will not lead to acceptably similar quantitative measurements between systems.

4.2 Cross-Calibration

Figures 10 and 11 show that in almost every case using the cross-calibrated iterations resulted in less of a difference in sphere measurements than did setting the iterations equal between systems. The only exceptions to this trend were at high (>100) iterations, wherein setting iterations equal resulted in slightly less different sphere measurements than the algorithm. For the most part, the hot rod phantom performed slightly better than the cold rod phantom, with substantially better performance for some TOF images. Size dependence seemed to be an issue, with the non-TOF to TOF cross-calibration resulting in larger abdomen phantom sphere differences compared to the non-TOF to non-TOF cross-calibration.

The data suggests that cross-calibration through post-smoothing control works well, with high iterations performing slightly better. Interestingly, post-smoothing control at clinical iterations performs better with the noisy images than high iterations. Size dependence seemed to be an issue here as well, with the spheres inside the Jaszczak phantom correlating better than the spheres inside the abdomen phantom. This was especially true with TOF images.

The poor non-TOF to TOF correlation between the Jaszczak phantoms and the spheres inside the abdomen phantom prompted further investigation. To test if the size dependence of PET was responsible, six hot spheres were placed inside the Jaszczak phantom container along with the hot rod Jaszczak insert. This equalized the object size between the Jaszczak insert and the spheres. Theoretically this should provide approximately equal TOF improvement since both the insert and spheres are hot regions. The very good TOF to non-TOF correlation of the Jaszczak insert with the spheres placed inside the Jaszczak phantom suggests that the poor TOF to non-TOF correlation with spheres placed in the abdomen phantom is indeed due to the known size dependence of PET. This is consistent with previous published works. For instance, a variable body size study by Wilson and Turkington showed that not only did TOF exhibit higher sphere contrast relative to non-TOF at every body size, but also that non-TOF sphere contrast converged in a more body size dependent way than TOF. (Wilson

and Turkington, TOF-PET Small-Lesion Image Quality Measured Over a Range of Phantom Sizes 2013)

Perhaps a correction factor can be derived to allow the Jaszczak phantom to correlate well with the abdomen phantom for non-TOF to TOF cross-calibrations. If this correction factor can be achieved, then the Jaszczak phantom would be much more practical as a tool for quantitation equalization than the abdomen phantom, due to the aforementioned filling and reproducibility difficulties associated with the abdomen phantom.

This would solve one major issue with the Jaszczak phantom. The two major potential issues remaining with the Jaszczak phantom would then be misalignment of the PET and CT images and axial resolution sensitivity. For the first major issue, any misalignment would be easy to detect and correct. To evaluate the second issue, more study needs to be done on the cross-calibration performance of the Jaszczak phantom when the systems being cross-calibrated possess larger axial resolution differences. If there are issues with the Jaszczak phantom in this case, then it needs to be seen if a phantom sensitive to axial resolution, such as the Defrise disk phantom, may be used to first fix any issues arising due to axial resolution differences.

Since the Jaszczak phantom works very well in standardizing small sphere measurements with the spheres in the Jaszczak phantom, this suggests that the Jaszczak

phantom would work well for cross-calibrating small lesion measurements in the skull due to the size similarities between the average skull and the Jaszczak phantom.

5. Conclusions

Cross-calibrating two non-TOF systems with iteration control using either Jaszczak phantom results in equivalent (<2%) differences in sphere measurements in the abdomen phantom and acceptable (<5%) differences in sphere measurements in the Jaszczak phantoms. Cross-calibrating a non-TOF system with a TOF system using this same method results in unacceptable sphere measurement differences in the abdomen phantom, and acceptable differences in the Jaszczak phantom.

Using post-smoothing control to cross-calibrate two non-TOF systems resulted in acceptable abdomen phantom sphere measurement differences and equivalent Jaszczak phantom sphere measurements at both clinical and high IxSS levels. Performing this same method between a non-TOF system and a TOF system resulted in unacceptable differences in abdomen phantom sphere measurements and acceptable differences in the Jaszczak phantom spheres at both IxSS levels.

With noisier image sets, iteration control resulted in unacceptable abdomen sphere measurement differences and acceptable Jaszczak sphere measurement differences when cross-calibrating two non-TOF systems. Cross-calibrating a non-TOF system with a TOF system using this method resulted in acceptably different abdomen phantom sphere measurements and acceptably different Jaszczak phantom sphere measurements in the IxSS range used at Duke for SBMT scans (30-42 IxSS). At 84 DSTE IxSS, the differences in Jaszczak phantom sphere measurements became more

borderline, with all short scan durations except one falling within the acceptable threshold.

Post-smoothing the noisy images from two non-TOF systems resulted in acceptable differences in abdomen phantom sphere measurements at clinical IxSS levels and equivalent sphere measurements in the Jaszczak phantom. High IxSS levels resulted in unacceptable abdomen and Jaszczak phantom sphere measurement differences. Cross-calibrating a non-TOF system with a TOF system using this method resulted in only the Jaszczak phantom sphere measurements at clinical IxSS being acceptable. The other three sphere measurements were unacceptable.

Overall, post-smoothing seems to be the best cross-calibration method for two reasons. First, it was more successful than iteration control in general at minimizing differences in sphere measurements. The second reason is more practical. It is difficult to actually achieve the exact IxSS predicted by the iteration control cross-calibration without being flexible with the number of iterations the reconstruction performs. This would not be acceptable clinically, as even images reconstructed with three iterations take a long time to reconstruct on the systems used clinically at Duke. In contrast, achieving the post-smoothing dictated by the cross-calibration is trivially easy and does not affect reconstruction time.

References

Adams, Michael C, Timothy G Turkington, Joshua M Wilson, and Terence Z Wong. "A Systematic Review of the Factors Affecting Accuracy of SUV Measurements." *American Journal of Roentgenology*, no. 195 (2010): 310-320.

Ben-haim, S, and P Ell. "18F-FDG PET and PET/CT in the evaluation of cancer treatment response." *The Journal of Nuclear Medicine*, no. 50 (January 2009): 88-99.

Boellaard, Ronald, Nanda C Krak, Otto S Hockstra, and Adriaan A Lammertsma. "Effects of Noise, Image Resolution, and ROI Definition on the Accuracy of Standard Uptake Values: A Simulation Study." *The Journal of Nuclear Medicine*, no. 45 (2004): 1519-1527.

Budinger, Thomas F. "Time-of-Flight Positron Emission Tomography: Status Relative to Conventional PET." *The Journal of Nuclear Medicine* 24, no. 1 (1983): 73-78.

Data Spectrum Corporation. *DSC QC PET Phantom Products.* March 17, 2007. http://www.spect.com/products-pet.html (accessed April 5, 2014).

Hudson, H Malcolm, and Richard S Larkin. "Accelerated Image Reconstruction Using Ordered Subsets of Projection Data." *IEEE Transactions on Medical Imaging* 13, no. 4 (1994): 601-609.

Jaskowiak, Chris J, Jesus A Bianco, Scott B Perlman, and Jason P Fine. "Influence of Reconstruction Iterations on 18F-FDG PET/CT Standardized Uptake Values." *The Journal of Nuclear Medicine*, 2005: 424-428.

Makris, Nikolaos E, Marc C Huisman, Paul E Kinahan, Adriaan A Lammertsma, and Ronald Boellaard. "Evaluation of strategies towards harmonization of FDG PET/CT studies in multicentre trials: comparison of scanner validation phantoms and data analysis procedures." *European Journal of Nuclear Medicine and Molecular Imaging*, no. 40 (2013): 1507-1515.

Scheuermann, Joshua S, Janet R Saffer, Joel S Karp, Anthony M Levering, and Barry A Siegel. "Qualification of PET scanners for use in multicenter cancer clinical trials: The American College of Radiology Imaging Network experience ." *The Journal of Nuclear Medicine*, no. 50 (July 2009): 1187-1193.

Shankar, Lalitha K, John M Hoffman, and Steve, et al Bacharach. "Consensus recommendations for the use of 18F- FDG PET as an indicator of therapeutic

response in patients in National Cancer Institute Trials." *The Journal of Nuclear Medicine*, no. 47 (June 2006): 1059-1066.

Smith, T Jordan. "Time-of-Flight PET Compared to Increased Scan Time in Low-Contrast Regions." Masters Thesis, 2011.

Tong, S, A M Alessio, and P E Kinahan. "Noise and signal properties in PSF-based fully 3D PET image reconstruction: an experimental evaluation ." *Physics in Medicine and Biology*, no. 55 (2010): 1453-1473.

Weber, W A. "Assessing tumor response to therapy." *The Journal of Nuclear Medicine*, no. 50 (May 2009): 1S-10S.

Weber, Wolfgang A, and Hinrich Wieder. "Monitoring chemotherapy and radiotherapy of solid tumors." *European Journal of Nuclear Medicine and Molecular Imaging*, no. 33 (May 2006): 27-37.

Westerterp, Marinke, Jan Pruim, Wim Oyen, Otto Hoekstra, Anne Paans, and et al. "Quantification of FDG PET studies using standardised uptake values in multi-centre trials: effects of image reconstruction, resolution and ROI definition parameters." *European Journal of Nuclear Medicine* 34, no. 3 (2007): 392-404.

Wilson, Joshua M, and Timothy G Turkington. "Multisphere phantom and analysis algorithm for PET image quality assessment ." *PHYSICS IN MEDICINE AND BIOLOGY* , no. 53 (May 2008): 3267–3278 .

Wilson, Joshua M, and Timothy G Turkington. "TOF-PET Small-Lesion Image Quality Measured Over a Range of Phantom Sizes ." *IEEE Transactions on Nuclear Science* 60, no. 3 (2013): 1589-1595.